Genotype-phenotype association study via new multi-task learning model

Zhouyuan Huo
Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260, United States
E-mail: zhouyuan.huo@pitt.edu

Dinggang Shen
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
E-mail: dinggang_shen@med.unc.edu

Heng Huang*
Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260, United States
E-mail: heng.huang@pitt.edu

Research on the associations between genetic variations and imaging phenotypes is developing with the advance in high-throughput genotype and brain image techniques. Regression analysis of single nucleotide polymorphisms (SNPs) and imaging measures as quantitative traits (QTs) has been proposed to identify the quantitative trait loci (QTL) via multi-task learning models. Recent studies consider the interlinked structures within SNPs and imaging QTs through group lasso, e.g. $\ell_{2,1}$-norm, leading to better predictive results and insights of SNPs. However, group sparsity is not enough for representing the correlation between multiple tasks and $\ell_{2,1}$-norm regularization is not robust either. In this paper, we propose a new multi-task learning model to analyze the associations between SNPs and QTs. We suppose that low-rank structure is also beneficial to uncover the correlation between genetic variations and imaging phenotypes. Finally, we conduct regression analysis of SNPs and QTs. Experimental results show that our model is more accurate in prediction than compared methods and presents new insights of SNPs.

Keywords: Quantitative Trait Loci; Single Nucleotide Polymorphisms (SNPs); Quantitative Traits (QTs); Multi-Task Learning.

1. Introduction

Research on the associations between genetic variations and imaging phenotypes is developing with the advance in high-throughput genotype and brain image techniques.[1–4] Alzheimers Disease Neuroimaging Initiative (ADNI) provides a suitable dataset for genotype-phenotype study, however it is still challenging to find out whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), genetic factors such as single nucleotide polymorphisms (SNPs) can be

*Corresponding Author. This work was partially supported by U.S. NIH R01 AG049371, NSF IIS 1302675, IIS 1344152, DBI 1356628, IIS 1619308, IIS 1633753

combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's Disease (AD). Given these data, researchers did the association study between genetic variation and imaging measures as quantitative traits (QTs), which was shown to have increased statistical power and decreased sample size requirements.[5] Through the analysis of strong associations between SNPs and imaging phenotypes, we can also identify candidate genes or loci which are relevant to the biological etiology of the disease.[2]

Traditional association studies use univariate or multivariate methods to discover the associations between single nucleotide polymorphisms (SNPs) and imaging measures as quantitative traits (QTs).[6,7] However, these methods treat each regression of imaging phenotype as an independent task, thus the correlations between SNPs and QTs are lost in this model. To solve this problem, regression analysis of SNPs and QTs has been proposed to identify the quantitative trait loci (QTL) via multi-task learning models.[4,8] In multi-task learning model, multiple tasks are handled jointly and dependently. For example, by imposing the interlinked structures within SNPs and imaging QTs through group lasso, e.g. $\ell_{2,1}$-norm,[9,10] it leads to better predictive results and more insights of the SNPs.[4] This assumption is suitable for the fact that only a small fraction of SNPs are responsible for the imaging manifestations of complex diseases. However, there are two limitations. Firstly, group sparsity is not enough for representing the intrinsic correlation between SNPs and imaging QTs. Apart from group sparsity, we can also benefit from the low-rank structure of the coefficient. Secondly, although $\ell_{2,1}$-norm regularization is common for the group sparsity, it is sensible to outliers.[11] For example, the value of $\ell_{2,1}$-norm of matrix $[[100], [0], [0]]$ is larger than $[[1], [1], [1]]$, however, the first matrix is more sparse rather than the second one.

In this paper, we propose a new multi-task learning model to analyze the associations between SNPs and QTs. We suppose that low-rank structure is also beneficial to uncover the correlation between genetic variations and imaging phenotypes. This assumption is reasonable because different SNPs may have similar effect on the imaging phenotypes. For example, both APOE SNPs rs429358 and rs7412 are the strongest known genetic risk factors for Alzheimer's Disease. In order to make the feature selection robust to outliers, we propose to use capped $\ell_{2,1}$-norm regularization in place of $\ell_{2,1}$-norm. We conduct regression analysis of SNPs and QTs from ADNI, and the experimental results show that our model is more accurate in prediction than compared methods and it presents new insights of SNPs as well.

2. Data Description

We use the dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). One goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. These data are obtained from 818 participants. Further information about ADNI can be found at see www.adni-info.org.

We use the genotype data[12] of all non-Hispanic Caucasian participants from the ADNI Phase 1 cohort. They were genotyped using the Human 610-Quad BeadChip. Only SNPs which belong to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of 4/18/2011[13] were selected after the standard quality control (QC) and imputation steps. The QC criteria for the

SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. After that, the quality-controlled SNPs were imputed using the MaCH software[14] to estimate the missing genotypes in the second pre-processing step. In this paper, we use 3123 SNPs in total. While most of them might be irrelevant to AD, only a small fraction of them are risk factors for the disease and associated with imaging phenotypes. For example, gene APOE and TOMM40 are known to be the contributors to AD.

Two widely employed automated MRI analysis techniques were used to process and extract imaging phenotypes from scans of ADNI participants as previously described.[3] First, Voxel-Based Morphometry (VBM)[15] is performed to define global gray matter (GM) density maps and extract local GM density values for target regions. Second, automated parcellation via FreeSurfer V4[16] is conducted to define volumetric and cortical thickness values for regions of interest (ROIs) and to extract total intracranial volume (ICV). All these measures were adjusted for the baseline ICV using the regression weights derived from the healthy control (HC) participants. Further details are available in.[3] In this paper, we use 36 ROIs from VBM and 24 ROIs from FreeSurfer which are known to be related to AD. VBM measures and FreeSurfer measures are treated as QTs for identifying QTLs independently.

3. Proposed Method

In this section, we propose a new multi-task learning model to study the intrinsic associations between SNPs and imaging phenotypes. Throughout our paper, we use $X \in \mathbb{R}^{d\times n}$ to denote the SNP data of all the ADNI participants, and $Y \in \mathbb{R}^{c\times n}$ to denote the selected imaging phenotypes, where n is the number of participants, d is the number of SNPs and c denotes the number of selected imaging phenotypes or QTs. It is a standard regression problem to predict continuous quantities Y using SNPs data X as follows:

$$\min_{W\in\mathbb{R}^{d\times c}} \|W^T X - Y\|_F^2 \tag{1}$$

The learned weight matrix W shows the importance of each SNP to predict imaging phenotypes, e.g. W_i^j denotes the importance of i-th SNP to predict j-th imaging phenotype. There are mainly three drawbacks of using model (1) as the objective function to learn the coefficient matrix W. Firstly, it is easy to overfit if there is no regularization, and the learned W is hard to generalize to new data. Secondly, the learned coefficient matrix W is not sparse. It is intuitive that only a small fraction of SNPs should be relevant to imaging quantitative traits (QTs), thus sparsity of W is a nontrivial property. The last but not the least, the associations within SNPs or imaging phenotypes are overlooked. Coefficient matrix W should come from a specific domain, we can impose a structured regularization on W to represent the intrinsic associations within SNPs or imaging phenotypes. We usually use l_2-norm regularization to avoid overfitting, however, the last two problems are still not solved yet. To handle these issues, we can treat the regression of each column of Y (each quantitative trait (QT)) as a task, then we can use multi-task learning model to learn multiple tasks jointly. The original problem (1) can be represented as a multi-task problem as follows:

$$\min_{W=[W^1,\ldots,W^T]\in\mathbb{R}^{d\times c}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} \|(W^t)^T x_{i,t} - y_{i,t}\|_2^2 + \text{Reg}(W) \tag{2}$$

where $T = c$ (the number of tasks), $n_t = n, \forall t \in \{1, ..., T\}$ (the number of samples in task t). In task t, $x_{i,t} = X^i$, which is the column i of X; $y_{i,t} = Y_t^i$, which is the element of Y at the position of row t and column i; W^t denotes the column t of matrix W. Reg(W) is the regularization we impose on the multi-task learning problem, and it represents our assumption of the correlation between multiple tasks , *e.g.* low-rank or group sparsity.[17,18] In the following context, we propose to impose two new regularization terms in the multi-task problem to learn the associations between SNPs and imaging phenotypes, one for genetic association and the other one for quantitative trait loci (QTLs) identification.

3.1. *Capped Trace Norm Regularization for Genetic Association*

In multi-task learning, we assume that the regression tasks between SNPs and imaging phenotypes are correlated. Then we can benefit from learning multiple tasks jointly. Their correlation can be represented by imposing a structure on the coefficient matrix W. In this paper, we assume that matrix W has a low-rank subspace, which is widely used in many applications, such as recommendation system[19,20] and multi-task learning.[21,22] This assumption is also fit for the genome-phenotype associations, because multiple SNPs may have similar effects on the imaging phenotype. For example, both APOE SNPs rs429358 and rs7412 are the strongest known genetic risk factors for Alzheimer's Disease. The non-convex rank minimization regularization Reg$(W) =$ rank(W) is hard to optimize, for simplicity, trace norm is proposed as the best convex relaxation for the rank minimization regularization as follows[23] :

$$\text{Reg}(W) = \|W\|_* = \sum_{i=1}^{\min\{d,c\}} \sigma_i(W) \tag{3}$$

where σ_i is the singular value of matrix W. However, there is a big gap between rank minimization regularization and trace norm regularization. When some non-zero singular values of W changes, the value of trace norm also changes. In contrast, the rank of matrix W keeps constant. Besides, trace norm is also sensitive to outliers.

In this paper, we propose to use a tighter approximation of rank minimization than trace norm. Capped trace norm is more general than trace norm and it is represented as follows:

$$\text{Reg}(W) = \sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i(W), \varepsilon_1\} \tag{4}$$

where ε_1 works as a threshold. If ε_1 is large enough, for any i, we have $\sigma_i(W) < \varepsilon_1$, then it is equal to trace norm regularization. When we reduce the value of ε_1, where $\varepsilon_1 \in \Big(min\{\sigma_i(W)\}, max\{\sigma_i(W)\}\Big)$, it's obvious that those singular values larger than ε_1 will be ignored in the optimization. So, instead of minimizing the sum of all singular values in the trace norm regularization, we focus on minimizing these singular values less than ε_1 and ignore large singular values. Therefore, capped trace norm regularization is more robust to outliers.

3.2. *Capped $\ell_{2,1}$-Norm Regularization for QTLs Identification*

There are 3123 SNPs in our dataset, and only a fraction of them is relevant to specific imaging quantitative traits (QTs). Therefore, W should be structured sparse, where each row of W is treated

as a unit. If SNP i is not important, $W_i = \mathbf{0} \in \mathbb{R}^{1\times c}$. $\ell_{2,0}$-norm regularization, $\text{Reg}(W) = \|\mathbf{w}\|_0$, minimizes the number of non-zero elements, where $\mathbf{w} \in \mathbb{R}^{d\times 1}$ and $\mathbf{w}_i = \|W_i\|_2$. However, it is a non-convex problem and hard to optimize. Alternatively, we usually use $\ell_{2,1}$-norm regularization enforce the structured sparsity on the learned coefficient matrix W:[4,9]

$$\text{Reg}(W) = \|W\|_{2,1} = \sum_{i=1}^{d} \|W_i\|_2 = \|\mathbf{w}\|_1 \tag{5}$$

where W_i denotes the i-th row of matrix W. Each row of W is treated as a unit, and if SNP i is negligible, $W_i = \mathbf{0} \in \mathbb{R}^{1\times c}$. Although $\ell_{2,1}$-norm regularization works fine, there is gap between $\ell_{2,0}$-norm regularization and $\ell_{2,1}$-norm regularization. Increasing the value of non-zero elements in $\mathbf{w}$ does not affect the number of its non-zero elements $\|\mathbf{w}\|_0$; on the contrary, $\|\mathbf{w}\|_1$ will increase. In this paper, we propose to use capped $\ell_{2,1}$-norm regularization as an alternative to $\ell_{2,0}$-norm as follows:

$$\text{Reg}(W) = \sum_{i=1}^{d} \min\{\|W_i\|_2, \varepsilon_2\} \tag{6}$$

Capped $\ell_{2,1}$-norm regularization is a better approximation of $\ell_{2,0}$-norm than $\ell_{2,1}$-norm. It treats $\|W_i\|_2$ equally if it is larger than ε_2, hence capped $\ell_{2,1}$-norm regularization is more robust to outliers. When ε_2 is large enough, we have $\min\{\|W_i\|_2, \varepsilon_2\} = \|W_i\|_2, \forall i$, thus capped $\ell_{2,1}$-norm is equal to $\ell_{2,1}$-norm.

To sum up, combining capped trace norm regularization and capped $\ell_{2,1}$-norm together makes our proposed objective function for multi-task learning (7) as follows:

$$\min_{W\in\mathbb{R}^{d\times c}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} \min \|(W^t)^T x_{i,t} - y_{i,t}\|_2^2 + \gamma_1 \sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i(W), \varepsilon_1\} + \gamma_2 \sum_{i=1}^{d} \min\{\|W_i\|_2, \varepsilon_2\} \tag{7}$$

where the notations are similar to problem (2). γ_1 and γ_2 are to balance the importance of two regularizations. In following sections, we will propose an efficient optimization algorithm for problem (7) and prove that it is sequence convergent.

4. Optimization Algorithm

In this section, we propose an efficient optimization algorithm to solve problem (7). Optimizing the non-smooth and non-convex problem (7) directly is very hard. Through re-weighted algorithm,[24] in each step, we can transform our objective function to a smooth and convex relaxed problem, so that we are able to compute the optimal solution to the new relaxed problem until convergence.

Firstly, we do Singular Value Decomposition (SVD) on the coefficient matrix W and we have $W = U\Sigma V^T$, where singular values $\sigma_i(W)$ of matrix W are in ascending order. Assuming there are k singular values smaller than ε_1, we define $D = \frac{1}{2}\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T$ where U^i is the i_{th} column of matrix U. Therefore, the second term in (7) can be represented as $\gamma_1 \text{Tr}(W^T D W)$. Secondly, we compute Z_{ii} for each row of matrix W:

$$Z_{ii} = \begin{cases} \frac{1}{2\|W_i\|_2} & \text{if } \|W_i\|_2 < \varepsilon_2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

All the non-diagonal elements of matrix Z are 0. Therefore, the third term in (7) can be represented by $\gamma_2 \text{Tr}(W^T Z W)$. When we fix the values of D and Z, the objective function (7) can be written as a smooth and convex problem as follows:

$$\min_{W=[W^1,\ldots,W^T]} \|W^T X - Y\|_F^2 + \gamma_1 \text{Tr}(W^T D W) + \gamma_2 \text{Tr}(W^T Z W) \tag{9}$$

where the loss term is from $\sum_{t=1}^{T} \sum_{i=1}^{n_t} \|(W^t)^T x_{i,t} - y_{i,t}\|_2^2 = \|W^T X - Y\|_F^2$ as per the definition of our variables. Finally, taking the derivative of (9) in terms of W and setting it to zero, we can get the optimal solution to the problem (9) as follows:

$$W = (X X^T + \gamma_1 D + \gamma_2 Z)^{-1} X Y^T \tag{10}$$

To sum up, our proposed optimization algorithm is presented in Algorithm 1.

Algorithm 1 Algorithm to solve problem (7)

Input: Training data for multiple tasks $X \in \mathbb{R}^{d \times n}, Y \in \mathbb{R}^{c \times n}$
Output: $W \in \mathcal{R}^{d \times c}$.
Initialize W.
while not converge **do**
  Compute D and Z via (4) and (8).
  Fix D and Z, and compute matrix W via (10).
end while

5. Convergence Analysis

By optimizing our model with Algorithm 1, we can solve the non-smooth and non-convex objective function (7). In this section, we presents the convergence analysis of our proposed algorithm.

Theorem 1. *Through Algorithm 1, the values of objective function (7) are non-increasing monotonically, and it will converge to a local solution.*

In order to prove Theorem 1, we need the following Lemmas.

Lemma 1. *According to,[25] any two hermitian matrices $A, B \in R^{n \times n}$ satisfy the following inequality:*

$$\sum_{i=1}^{n} \sigma_i(A)\, \sigma_{n-i+1}(B) \leq \text{Tr}\left(A^T B\right) \leq \sum_{i=1}^{n} \sigma_i(A)\, \sigma_i(B) \tag{11}$$

where $\sigma_i(A)$, $\sigma_i(B)$ are singular values sorted in the same order.

Lemma 2. *Let $W = U \Sigma V^T$, Σ is a diagonal matrix and σ_i are singular values of W in ascending order. There are k singular values less than ε_1. $\hat{W}$ is coefficient matrix in next iteration by using Algorithm 1, and $\hat{W} = \hat{U} \hat{\Sigma} \hat{V}^T$, where $\hat{\sigma}_i$ are singular values of $\hat{W}$ in ascending order and U^i is the*

i-th column of U. There are $\hat{k}$ singular values less than ε_1. So it is true that:

$$\sum_{i=1}^{\min\{d,c\}} \min\{\hat{\sigma}_i, \varepsilon_1\} - \frac{1}{2}\operatorname{Tr}\Big(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T \hat{W}\hat{W}^T\Big) \tag{12}$$

$$\leq \sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i, \varepsilon_1\} - \frac{1}{2}\operatorname{Tr}\Big(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T W W^T\Big) \tag{13}$$

Proof: It's obvious that $\sigma_i - 2\hat{\sigma_i} + \sigma_i^{-1}{\hat{\sigma_i}}^2 = \frac{1}{\sigma_i}\left(\sigma_i^2 - 2\sigma_i\hat{\sigma_i} + {\hat{\sigma_i}}^2\right) \geq 0$. Thus we have:

$$\sum_{i=1}^{k}\left(\hat{\sigma}_i - \frac{1}{2}\sigma_i^{-1}\hat{\sigma}_i^2\right) \leq \frac{1}{2}\sum_{i=1}^{k}\sigma_i \tag{14}$$

Because there are $\hat{k}$ singular values of $\hat{W}$ less than ε_1 and they are sorted in ascending order, so first $\hat{k}$ singular values $\hat{\sigma}_i$ are less than ε_1. Therefore, no matter $\hat{k} \geq k$ or $\hat{k} < k$, it holds that:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i - \hat{k}\varepsilon_1 \leq \sum_{i=1}^{k}\hat{\sigma}_i - k\varepsilon_1 \tag{15}$$

Combining (14) and (15), we get the following inequality:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i - \frac{1}{2}\sum_{i=1}^{k}\sigma_i^{-1}\hat{\sigma}_i^2 - \hat{k}\varepsilon_1 \leq \frac{1}{2}\sum_{i=1}^{k}\sigma_i - k\varepsilon_1 \tag{16}$$

Suppose there are $n = \min\{d, c\}$ singular values in total, adding $n\varepsilon_2$ on both sides, we are able to get the following inequality:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i + \left(n - \hat{k}\right)\varepsilon_1 - \tfrac{1}{2}\sum_{i=1}^{k}\sigma_i^{-1}\hat{\sigma}_i^2 \leq \sum_{i=1}^{k}\sigma_i + (n-k)\,\varepsilon_1 - \tfrac{1}{2}\sum_{i=1}^{k}\sigma_i \tag{17}$$

According to the definition of matrix D in (4), the following equality holds that:

$$\frac{1}{2}\operatorname{Tr}(W^T D W) = \frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T W W^T\right) = \frac{1}{2}\operatorname{Tr}\left(U\Lambda U^T U\Sigma^2 U^T\right) = \frac{1}{2}\sum_{i=1}^{k}\sigma_i \tag{18}$$

where Λ is the diagonal matrix where its first k elements are $\sigma_i^{-1}, i \in \{1, ..., k\}$ and other elements are 0. Via Lemma 1, we have:

$$\tfrac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T\hat{W}\hat{W}^T\right) = \tfrac{1}{2}\operatorname{Tr}\left(U\Lambda U^T\hat{U}\hat{\Sigma}^2\hat{U}^T\right) \geq \tfrac{1}{2}\sum_{i=1}^{k}\sigma_i^{-1}\hat{\sigma}_i^2 \tag{19}$$

Substituting (18) and (19) in the inequality (17), it is satisfied that:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i + \left(n - \hat{k}\right)\varepsilon_1 - \frac{1}{2}\operatorname{Tr}\Big(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T\hat{W}\hat{W}^T\Big)$$

$$\leq \sum_{i=1}^{k}\sigma_i + \Big(n-k\Big)\varepsilon_1 - \frac{1}{2}\operatorname{Tr}\Big(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T W W^T\Big) \tag{20}$$

Finally, the following inequality holds that:

$$\begin{aligned}&\sum_{i=1}^{\min\{d,c\}}\min\{\hat{\sigma}_i,\varepsilon_1\}-\frac{1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T\hat{W}\hat{W}^T\Big)\\ \leq&\sum_{i=1}^{\min\{d,c\}}\min\{\sigma_i,\varepsilon_1\}-\frac{1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^TWW^T\Big)\end{aligned}\tag{21}$$

Lemma 3. *We define* $z=\begin{cases}\frac{1}{2|e|} & \textit{if } |e|<\varepsilon_2\\ 0 & \textit{otherwise}\end{cases}$, *then the inequality holds that* $\min\{|\hat{e}|,\varepsilon_2\}-z\hat{e}^2\leq\min\{|e|,\varepsilon_2\}-ze^2$.

Proof: If $|e|<\varepsilon_2$, we have $z=\frac{1}{2|e|}$. Via Lemma 2, let W and $\hat{W}$ be scalars $|e|$ and $|\hat{e}|$ respectively, thus $\sigma(|e|)=|e|$ and $\sigma(|\hat{e}|)=|\hat{e}|$. We substitute W, $\hat{W}$ and z in the inequality (21), it holds that:

$$\min\{|\hat{e}|,\varepsilon_2\}-z\hat{e}^2\leq\min\{|e|,\varepsilon_2\}-ze^2\tag{22}$$

On the other hand, if $|e|\geq\varepsilon_2$, we have $z=0$. The following inequality always holds:

$$\min\{|\hat{e}|,\varepsilon_2\}\leq\min\{|e|,\varepsilon_2\}\tag{23}$$

Right now, we are able to prove Theorem 1 by using Lemma 2 and Lemma 3 above.

Proof: According to the step 2 in Algorithm 1, matrix W denotes the current values of our model, after we obtain the analysis solution $\hat{W}$ of function (9) through (10) . Therefore, it is guaranteed that:

$$\begin{aligned}&\|\hat{W}^TX-Y\|_F^2+\gamma_1\mathrm{Tr}(\hat{W}^TD\hat{W})+\gamma_2\mathrm{Tr}(\hat{W}^TZ\hat{W})\\ \leq&\|W^TX-Y\|_F^2+\gamma_1\mathrm{Tr}(W^TDW)+\gamma_2\mathrm{Tr}(W^TZW)\end{aligned}\tag{24}$$

We define, $|e|=\|W_i\|_2$, $|\hat{e}|=\|\hat{W}_i\|_2$ and $z_i=Z_{ii}$. after substituting the value of $|e|$ in Lemma 3, we have:

$$\min\{\|\hat{W}_i\|_2,\varepsilon_2\}-Z_{ii}\|\hat{W}_i\|_2^2\leq\min\{\|W_i\|,\varepsilon_2\}-Z_{ii}\|W_i\|_2^2\tag{25}$$

By summing up from $i=1$ to d, and multiplying both sides with γ_2, then the following inequality holds that:

$$\gamma_2\sum_{i=1}^{d}\min\{\|\hat{W}_i\|_2,\varepsilon_2\}-\gamma_2\mathrm{Tr}(\hat{W}^TZ\hat{W})\leq\gamma_2\sum_{i=1}^{d}\min\{\|W_i\|_2,\varepsilon_2\}-\gamma_2\mathrm{Tr}(W^TZW)\tag{26}$$

where $\sum\limits_{i=1}^{d}Z_{ii}\|W_i\|_2^2=\mathrm{Tr}(W^TZW)$.

Via Lemma 2, we can easily know that:

$$\begin{aligned}&\gamma_1\sum_{i=1}^{\min\{d,c\}}\min\{\hat{\sigma}_i,\varepsilon_1\}-\frac{\gamma_1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T\hat{W}\hat{W}^T\Big)\\ \leq\gamma_1&\sum_{i=1}^{\min\{d,c\}}\min\{\sigma_i,\varepsilon_1\}-\frac{\gamma_1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^TWW^T\Big)\end{aligned}\tag{27}$$

Finally, we combine inequalities (18), (24), (26) and (27), then we know that the objective value sequence is monotonically non-increasing:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{n_t}\|(\hat{W}^t)^T x_{i,t}-y_{i,t}\|_2^2+\gamma_1\sum_{i=1}^{\min\{d,c\}}\min\{\sigma_i(\hat{W}),\varepsilon_1\}+\gamma_2\sum_{i=1}^{d}\min\{\|\hat{W}_i\|_2,\varepsilon_2\}\\
\leq&\sum_{t=1}^{T}\sum_{i=1}^{n_t}\|(W^t)^T x_{i,t}-y_{i,t}\|_2^2+\gamma_1\sum_{i=1}^{\min\{d,c\}}\min\{\sigma_i(W),\varepsilon_1\}+\gamma_2\sum_{i=1}^{d}\min\{\|W_i\|_2,\varepsilon_2\} \qquad (28)
\end{aligned}$$

After several iterations, $\hat{W}\approx W$, the derivative of the objective function (9) is close to zero. So far, it is clear that the values of our proposed objective function will not increase by using our optimization algorithm, so we prove Theorem 1 that our optimization algorithm is non-increasing monotonically. We also know that the objective function (7) is lower bounded. We can conclude that our optimization algorithm is sequence convergent.

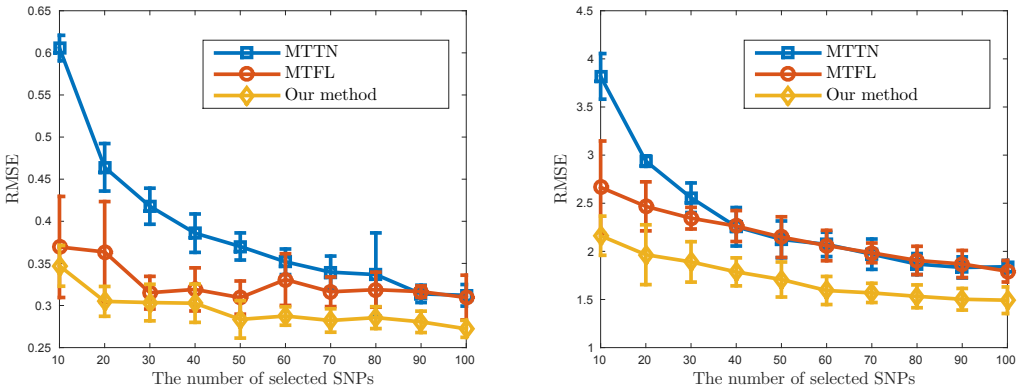
6. Experimental Results and Discussions

In this section, we evaluated our proposed model with other multi-task learning methods. The experimental dataset is from the ADNI cohort. Our goal is to select a subset of SNPs to predict the imaging phenotypes accurately. We conduct our experiments on two imaging phenotypes, FreeSurfer and VBM separately. There are two compared methods, multi-task learning with joint feature selection (MTFL)[9] and multi-task learning with trace norm regularization (MTTN),[26] both of them use least square loss to do regression. It is easy to observe that MTFL and MTTN can be represented by our proposed model. If $\gamma_2=0$ and $\varepsilon_1=\infty$, it is MTFL; if $\gamma_1=0$ and $\varepsilon_2=\infty$, it is MTTN.

We conduct 5-fold cross-validation, where 4 folds are training data and 1-fold is testing data. Then we perform internal 5-fold cross-validation on the training data, and tune parameters γ_1 and γ_2 in the range of $\{10^{-4},10^{-3},...,10^{3},10^{4}\}$. Through the learned coefficient matrix W, we compute the weight of i_{th} SNP over all tasks by using $\sum_{j=1}^{c}|W_i^j|$. Then, we pick up the top $\{10,20,...,90,100\}$ SNPs to predict the regression responses of the testing data. For our method, although there are two other parameters ε_1 and ε_2 in the objective function (7), their values are set automatically during the optimization. In the first 5 iterations, ε_1 is set to be the 5_{th} largest singular value in $\sigma_i(W)$ and ε_2 is set to bet the 5_{th} largest value of SNP weight $\|W_i\|_2$. After that, we fix the values of ε_1 and ε_2 until convergence. In our experiments, we always stop our algorithm 1 after 20 iterations. The performance of compared method is evaluated by Root Mean Square Error (RMSE), which is a widely used measurement for regression analysis.

6.1. *Improved Phenotype Prediction*

The experimental results are presented in Figure 1. It shows the mean and standard deviation of the RMSEs obtained from 5 trails. In Figure 1, we observe that our proposed method consistently outperforms other two compared methods in both VBM phenotypes and FreeSurfer phenotypes. When we change the number of selected SNPs in our experiments, we can find out that models with joint feature selection regularization, $\ell_{2,1}$-norm or capped $\ell_{2,1}$-norm, are more stable. On the contrary, MTTN is very sensitive to the number of selected SNPs, and its performance is far worse when the number of SNPs is small. We can also observe that when the number of selected SNPs is larger than



(a) VBM (b) FreeSurfer

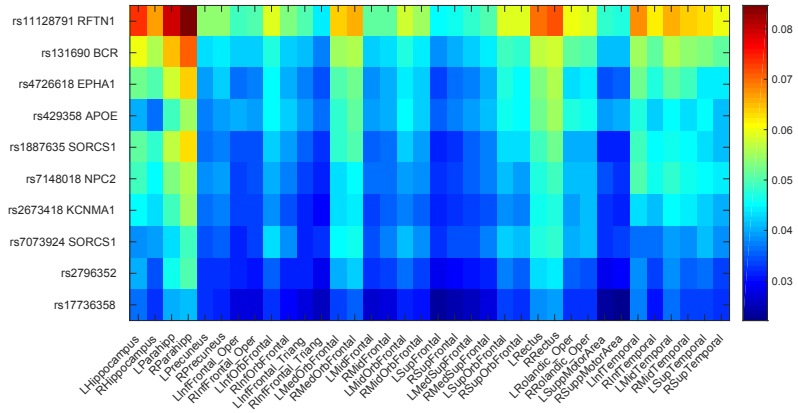
Fig. 1. Experimental results of three compared methods on two phenotypes. Average values are taken from five cross-validation and each error bar denotes $\pm$ standard deviation. Figure 1(a) shows the results of VBM phenotypes, Figure 1(b) shows the results of Freesurer phenotypes.

50, the improvement of prediction is small. Thus, we can draw a conclusion that our assumption of sparsity of coefficient matrix is correct. Although there are 3123 SNPs in our experiment, only a fraction of them is responsible for the imaging phenotypes.

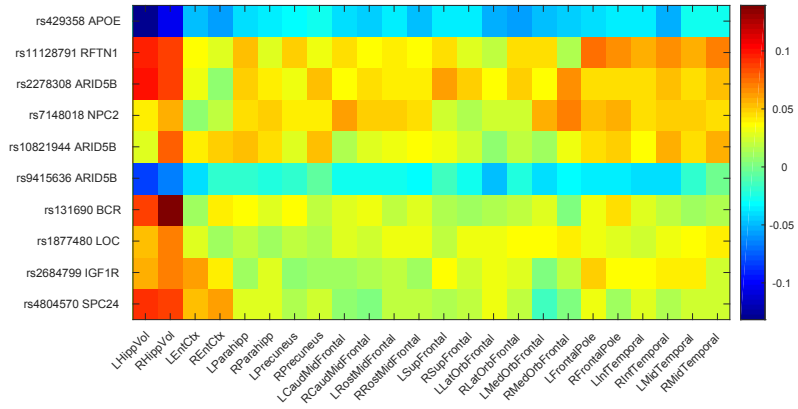
We also conduct ablation study of our method by setting $\gamma_1 = 0$ or $\gamma_2 = 0$ respectively. Table 1 presents the performance of compared methods when we select 20, 40 and 60 SNPs to predict imaging phenotypes. Firstly, we set $\gamma_2 = 0$, and our model becomes least square loss with capped trace norm regularization. We compare this model with MTTN, and experimental results demonstrate the effectiveness of capped trace norm. We also set $\gamma_1 = 0$, and our model is least square loss with capped $\ell_{2,1}$-norm regularization. We compare this model with MTFL, and it is clear that our method is more accurate in the prediction of imaging phenotypes. When we combine both of these two terms, $\gamma_1 \neq 0$ and $\gamma_2 \neq 0$, our model obtain the best results. We can draw a conclusion that although the performance of our method when $\gamma_2 = 0$ is much worse than the performance when $\gamma_1 = 0$, imposing low-rank structure on coefficient matrix is still beneficial to the regression analysis. Therefore, it is consistent with the fact that multiple SNPs may have similar effects on the imaging phenotypes.

Table 1. Ablation study of our method measured by RMSE. Value: RMSE, (comparison with corresponding method), e.g RMSE of capped $\ell_{2,1}$-norm (RMSE of capped $\ell_{2,1}$-norm $-$ RMSE of MTFL)

Phenotype	Method	20	40	60
VBM	capped trace norm ($\gamma_2 = 0$)	0.4566 (-0.0075)	0.3754 (-0.0105)	0.3398(-0.0120)
	capped $\ell_{2,1}$ ($\gamma_1 = 0$)	0.3381 (-0.0255)	0.3124 (-0.0067)	0.3066(-0.0242)
	Our Method	**0.3049**	**0.3027**	**0.2875**
FreeSurfer	capped trace norm ($\gamma_2 = 0$)	2.8623 (-0.0756)	2.2043(-0.0511)	1.9677 (-0.1047)
	capped $\ell_{2,1}$ ($\gamma_1 = 0$)	2.2030 (-0.2646)	1.8747 (-0.3883)	1.6389 (-0.4215)
	Our Method	**1.9653**	**1.7869**	**1.5934**



(a) VBM



(b) FreeSurfer

Fig. 2. Heat maps of regression coefficients learned genetic variations and quantitative traits (QTs). Top 10 selected SNPs of each matrix are visualized. Figure 2(a) shows the results from the regression of VBM measures, Figure 2(b) shows the results from the regression of FreeSurfer measures.

6.2. *Gene Selection*

Figure 2 visualizes the coefficient of top selected 10 SNPs. APOE is known to have relationship with the Alzheimer's disease (AD). Similar to previous research,[3,8] we find that APOE rs429358 shows the strongest associations with all imaging quantitative traits (QTs), especially in Figure 2(b). Clearly, our propose model is able to identify important quantitative trait loci (QTL) via joint regression analysis. Besides, we also observe that RFTN1 rs11128791 also takes important role in the imaging phenotypes, which is not identified in previous methods. These newly identified SNPs are highly correlated with the imaging phenotypes which are related to AD. They all have potential to serve as a useful generic risk factor for AD.

7. Conclusion

In this paper, we propose a new multi-task learning model with capped trace norm and capped $\ell_{2,1}$-norm regularizations. Capped trace norm helps to discover intrinsic structures within SNPs and imaging phenotypes; capped $\ell_{2,1}$-norm is more robust to select important SNPs. We propose efficient

algorithm to solve our model and provide convergence analysis. Finally, we conduct experiments on genotype-phenotype dataset from ADNI. Experimental results show that (1) our model works better in imaging phenotype prediction and (2) it helps to identify important quantitative trait loci (QTLs), which would be useful for the investigation of the generic risk factor for AD.

References

1. C. E. Bearden, T. G. Van Erp, R. A. Dutton, H. Tran, L. Zimmermann, D. Sun, J. A. Geaga, T. J. Simon, D. C. Glahn, T. D. Cannon *et al.*, *Cerebral Cortex* **17**, 1889 (2006).
2. S. G. Potkin, G. Guffanti, A. Lakatos, J. A. Turner, F. Kruggel, J. H. Fallon, A. J. Saykin, A. Orro, S. Lupoli, E. Salvi *et al.*, *PloS one* **4**, p. e6501 (2009).
3. L. Shen, S. Kim *et al.*, *Neuroimage* **53**, 1051 (2010).
4. H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin and L. Shen, *Bioinformatics* **28**, 229 (2012).
5. S. G. Potkin, J. A. Turner *et al.*, *Cogn Neuropsychiatry* **14**, 391 (2009).
6. D. H. Ballard, J. Cho and H. Zhao, *Genetic epidemiology* **34**, 201 (2010).
7. J. Bralten, A. Arias-Vásquez, R. Makkinje, J. A. Veltman, H. G. Brunner, G. Fernández, M. Rijpkema and B. Franke, *American Journal of Psychiatry* **168**, 1083 (2011).
8. H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen and A. D. N. Initiative, *Bioinformatics* **28**, i127 (2012).
9. A. Argyriou, T. Evgeniou and M. Pontil, *Machine Learning* **73**, 243 (2008).
10. G. Obozinski, B. Taskar and M. Jordan, *Statistics Department, UC Berkeley, Tech. Rep* **2** (2006).
11. H. Gao, F. Nie, W. Cai and H. Huang, Robust capped norm nonnegative matrix factorization: Capped norm nmf, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.
12. A. J. Saykin, L. Shen *et al.*, *Alzheimers Dement* **6**, 265 (2010).
13. L. Bertram, M. B. McQueen *et al.*, *Nat Genet* **39**, 17 (2007).
14. Y. Li, C. J. Willer *et al.*, *Genet Epidemiol* **34**, 816 (2010).
15. J. Ashburner and K. J. Friston, *Neuroimage* **11**, 805 (2000).
16. B. Fischl, D. H. Salat *et al.*, *Neuron* **33**, 341 (2002).
17. J. Zhou, J. Chen and J. Ye, Multi-task learning: Theory, algorithms, and applications, in *URL https://www. siam. org/meetings/sdm12/zhou_chen_ye. pdf*, 2012.
18. Z. Huo, D. Shen and H. Huang, New multi-task learning model to predict alzheimer's disease cognitive assessment, in *Medical Image Computing and Computer-Assisted Intervention*, 2016.
19. C.-J. Hsieh and P. Olsen, Nuclear norm minimization via active subspace selection, in *International Conference on Machine Learning*, 2014.
20. Z. Huo, J. Liu and H. Huang, Optimal discrete matrix completion, in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.
21. T. K. Pong, P. Tseng, S. Ji and J. Ye, *SIAM Journal on Optimization* **20**, 3465 (2010).
22. Z. Huo, F. Nie and H. Huang, Robust and effective metric learning using capped trace norm: Metric learning via capped trace norm, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
23. E. J. Candès and T. Tao, *IEEE Transactions on Information Theory* **56**, 2053 (2010).
24. F. Nie, J. Yuan and H. Huang, Optimal mean robust principal component analysis, in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014.
25. C. Theobald, An inequality for the trace of the product of two symmetric matrices, in *Mathematical Proceedings of the Cambridge Philosophical Society*, (02)1975.
26. S. Ji and J. Ye, An accelerated gradient method for trace norm minimization, in *Proceedings of the 26th annual international conference on machine learning*, 2009.